DOCUMENT RESUME

ED 104 937 TM 004 395

AUTHOR Rowley, Glenn
TITLE A Rationale for Assessing the Reliability of an Observational Measure.
PUB DATE [Apr 75]
NOTE 35p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, D.C., March 30-April 3, 1975)

EDRS PRICE MF-$0.76 HC-$1.95 PLUS POSTAGE
DESCRIPTORS *Behavior; *Classroom Observation Techniques; Correlation; Evaluation Methods; Measurement Techniques; *Observation; Statistical Analysis; *Test Reliability

ABSTRACT

The use of the intraclass correlation in determining reliability is discussed and shown to be both appropriate and simple to use in the case of an observational measure, provided that observations are made on at least two occasions. The interpretation of such coefficients is explained in terms of generalizability theory, and real data are used to demonstrate how such coefficients can be interpreted and computed. Finally, an empirical study is described which investigates the effect on reliability of varying the number and length of the observation periods. (Author)

ED104937

A RATIONALE FOR ASSESSING THE RELIABILITY OF AN OBSERVATIONAL MEASURE

Glenn Rowley

The Ontario Institute for Studies in Education

U.S. DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Presented at the Annual Meeting of the American Educational Research Association

Washington, D. C.

March 30 - April 4, 1975

TM 004 395

Researchers in the area of classroom observation have been greatly troubled by questions concerning the reliability of the measures that they obtain. Until recently, this concern was frequently assuaged by the routine computation of one or more coefficients of observer agreement (see Frick and Semmel, 1974). However, the work of Medley and Mitzel (1963) and McGaw, Wardrop and Bunda (1972) has clearly established the inadequacy of observer agreements alone as indices of reliability. The variance components approach which they propose enables the researcher to pinpoint multiple sources of error, and to compute a number of different reliability coefficients for different purposes.

Unfortunately, the literature does not indicate that these methods have gained wide acceptance, at least in practice. The most likely reason for this would appear to be the implication in both papers (or at least read into them) that the estimation of reliability properly requires a fully-fledged reliability study, using multiple observers fully crossed with classrooms, and (following McGaw et al., 1972) crossed also with situations. The magnitude of such a study is far beyond the resources of most researchers, nor does such an undertaking relate very closely to the purposes of their own studies (typically, to make some statement about teacher or pupil behavior, and possibly its relationship with educational outcomes). Consequently, a common practice has been to avoid the question of reliability altogether, or else to report a coefficient of observer agreement, knowing full well its inadequacy for that purpose.

Reliability and observer agreement

Clearly, if different observers cannot agree in coding the same events, the observation system will not yield reliable results. For this reason,

observer agreement would seem to be a logical and useful matter to investigate during the development of an observational system and the refinement of its categories. If category definition is unclear, or even ambiguous, this could be expected to result in low agreement between observers. If only particular categories are ill-defined, or overlap, then coefficients based on those categories should indicate so. Thus the computation of coefficients of observer agreement ought to be of use to the investigator at least in the early stages of the development of an instrument.

Unfortunately, the use of measures of observer agreement has become confused with reliability. Inter-observer agreement is no guarantee of reliability, yet, in classroom observational literature, even writers of stature have confused these two distinct concepts; e.g. Rosenshine and Furst (1973) wrote: "Observer agreement is the most common form of reliability (p. 168)". This statement is, in fact, quite misleading, since observer agreement does not measure reliability in the usual sense of the word. This has been pointed out by many writers, including Brown, Mendenhall and Beaver (1968, p. 4), Medley and Mitzel (1963, p. 310) and, most succintly, by Westbury (1967):

> In interaction studies, reliability is usually defined simply as the agreement between two observers working over the same data. The establishment of consistency between observers is of course, crucial, but the domination of this concept of reliability denies some serious theoretical and statistical problems, and flies in the face of evidence, reported in passing in several studies, which suggests that teacher behaviour requires more than a few periods of observation to secure a representative summary (pp. 125-126).

McGaw et al. (1972) commented on the lack of impact of such discussions on the observational literature, suggesting (p. 16) that the confusion has arisen from a failure to distinguish primacy from prime importance. While observer agreement is one of the first issues to be faced by the developer of an

observational system, it is not the most important. In fact, if the reliability of an observational measure proves to be low, even a high measure of inter-observer agreement is of no account. A simple illustration may serve to demonstrate this point. Suppose, in a category system, we were to progressively reduce the number of categories by amalgamating them. As the number of categories gets smaller, the precision of measurement is lessened, but the percent of observer agreement could normally be expected to increase. In the limit of course, we reach the ultimate category system, having one hundred percent observer agreement, zero reliability, and only one category!

In an attempt to clear away the confusion surrounding terminology, Medley and Mitzel (1963) suggested the following three definitions:

> We will use the term reliability coefficient to refer to the correlation to be expected between scores based on observations made by different observers at different times. The correlation between scores based on observations made by different observers at the same time will be referred to as a coefficient of observer agreement. A correlation between scores based on observations made by the same observer at different times will be referred to as a stability coefficient (pp. 253-254).

These definitions were based on Medley and Mitzel's (1958) paper, and allowed for but one reliability coefficient, specifying in effect just which sources of variance were to be considered as "error" in estimating this coefficient. In particular the definition implies that variation in behavior from one occasion to the next be regarded as error.

McGaw et al. (1972), presenting their arguments in terms of generalizability theory (Cronbach, Gleser, Nanda, and Rajaratnam, 1972), took issue with Medley and Metzel (1963) on one point: namely, their assumption "that instability of behavior over occasions (i.e., time) is due to random error in one or both of the environment or the person (object) (p. 16)." This implies, they argued,

that there can be no lawful change in the characteristic being measured; an assumption which, in the case of teaching behavior, they hold untenable. Hence they presented an approach which was essentially similar to that of Medley and Mitzel (1963), but which allows for situation effects, and considers neither systematic changes in behavior over situations, nor systematic differences among teachers in their changes in behavior (teachers by situations interaction) as contributing to error. The approach advocated by McGaw et al. (1972) treats situations as a factor crossed with teachers, but occasions as a factor nested within teachers and situations. This would be appropriate in a case where a number of teachers were instructed to teach a number of lessons each of two or more types, and the investigator was interested in detecting differences both among teachers, and among the types of lesson. (Here the factor "situations" is identified with the type of lesson -- lecture, discussion, laboratory, etc. -- although it could equally well be used to refer to other characteristics of the lesson -- subject matter, grade level, etc.)

It is of interest that the approach of McGaw et al., like that of Medley and Mitzel, has been cited rather more frequently than it has been used. In fact no instance has come to light in which either approach has been used in an empirical study, although Medley and Mitzel (1963, p. 315) used data from an unpublished study to illustrate the procedures they had described. They found that

> Variation from situation to situation within the same class ... appears greater than variation in average behavior from one class to another (p. 316).

and concluded that

> In order to measure differences between classes reliably, therefore, it is necessary to observe each class in a number of situations, so that the fluctuations can cancel one another out (p. 317).

Just how much observation is necessary in order to achieve some kind of stability is not clear. However, it seems apparent that the few hours of observation used in most studies (frequently in large blocks) ar insufficient. As Westbury (1967) observed, this becomes a major problem whenever it is intended to use a sample of teaching behavior as a basis for an inference about a relationship between teaching behaviour and some learning outcomes. It first needs to be established that the observed sample of behavior constitutes a valid basis for generalizing to the universe of interest, which will almost certainly be the teachers' behavior over a substantially longer period of time.

Defining reliability

When we speak of the reliability of a test, it is usually fairly clear that the term "reliability" refers to the scores obtained by some sample of examinees on that test. Because a single test is typically used to produce a single score, we rarely have to ask: "The reliability of what?". If a single test were used to produce a number of different scores (e.g., if it contained two or more subtests), we would not want to speak of the reliability of the test itself, but rather of the reliabilities of each of the subtests. We would be aware, also, that any reliability coefficient that we might compute would depend to some extent on factors other than the test itself: the group of examinees, their range of ability, their motivation, the conditions of administration, including the time limit (if any), and the sources of error which are taken into account by the particular reliability estimate being used (see Stanley, 1971, for details).

Unfortunately, what may be obvious in one context need not be at all obvious in another. In the context of classroom observational research, it has frequently been asserted that reliability is a desirable property; it has not always been clear just what it is that is supposed to possess this attribute. In order to clarify the discussion which follows, the following definitions are offered:

> An observational instrument is a set of procedures by means of which an observer can record and categorize the behavior of a subject or subjects. It normally consists of a number of items, to which the observer responds in some way dependent on the behavior he has observed.
>
> An observational record is a set of data (usually in the form of symbols) which describes the behavior of one or more subjects during one or more periods of observation.
>
> An observational measure is a procedure for using an observational record to assign scores to each of the subjects of observation; each score so assigned being assumed to reflect some characteristic of the behavior of that subject.

By way of example, Flanders' Interaction Analysis (1970) would be described as an instrument; when it is used to observe teachers, and the data are recorded, we have an observational record; and when these data are used to compute an indirect/direct ratio (or any other score) for each teacher, we have an observational measure.

Workers in observational research have found reliability to be a troublesome question. Rosenshine (1971), described the traditional concept of reliability (the ability of the measure to distinguish between individuals)

as "particularly intriguing (p. 21)", and Brown et al. (1968), referred to it as "a tricky concept", commenting that "although everybody in educational research reads reliability coefficients, few seem to really understand (or care) what these mean or how they were obtained (p. 3)." No doubt there are many reasons for this confusion. One would seem to be that writers in the area have not made sufficiently clear to their readers that reliability is a property of a measure (in the sense defined above), and not of an instrument, or of a record. It needs to be established that an instrument itself is neither reliable nor unreliable -- it is only when the instrument has been used to collect data, and when the data have been manipulated in some way to produce scores, that we can speak sensibly about reliability. A single instrument can produce scores which are reliable, and other scores which are unreliable. Even one measure may be reliable or unreliable, depending on the manner in which the instrument is used, the subjects observed, the skill of the observer, and the number and length of observation periods. And yet even the most informed writers on the subject use phrases like "the reliability of observations of teachers' classroom behavior" (Medley and Mitzel, 1958), "the reliability of observation schedules" (McGaw et al. 1972, p. 13), "reliabilities of observational records" (Frick and Semmell, 1974, p. 1). It seems worth noting at the outset that the discussion to follow is concerned with the reliability of observational measures, and that reliability of any given measure will be dependent on a host of factors other than just the instrument by means of which or which it was obtained. In passing, it may be noted that one of the most valuable things to know about an instrument would be which measures produced from it are reliable and which are not, and under what conditions.

Suppose then, that an instrument is used to observe teachers, and a measure X is obtained. Then each teacher will have a score x on that measure, and we may speak sensibly of the variance σ_x^2 of those observed scores. It has been traditional, both in psychometrics (Lord and Novick, 1968, p. 61) and in observational literature (Medley and Mitzel, 1963, p. 309) to define the reliability ρ_{xx} as

$$\rho_{xx} = \frac{\sigma_t^2}{\sigma_x^2} ,$$

where σ_t^2 is the variance of true scores. In psychometrics, the definition of σ_x^2 has been less of a problem. The real difficulty has been in estimating σ_t^2, since this is an unobservable quantity.

With tests, we may conceive of a true score in its Platonic sense (see Lord and Novick, 1968, pp. 39-44, for details). Such a conception of true score has little relevance to observational measures, where the characteristic being measured typically has no existence (either real or hypothesized) except insofar as it is manifested in the subject's behavior. Consequently, rather than ask "what is the score which this person truly deserves?", we ask "if we were able to observe all of the relevant behavior of this person, what score would we then assign to him?". This leads to a definition of true score which is essentially similar to that used in psychometrics: the expected value, over repeated observations, of observed score.

In discussing the reliability of observational measures, Medley and Mitzel (1963) commented:

> There should be no controversy about the definition of σ_t^2 in this case. Suppose that the total score of class c on i items, based on records made by teams of r recorders on s visits, is x_{cris}. If the scale is supposed to measure differences between classes, and if any and all idiosyncrasies of observers, items, or

> situations are regarded as sources of distortion or error, then the true score of class c would be the mean of all the scores class c could get with any possible combination of i items, r recorders, and s situations equivalent to the i items, r recorders, and s situations actually used. The symbol σ_t^2 will be used to represent the variance of these true scores about the mean of all such true scores in the population of classes represented by the c classes actually visited.
>
> The definition of σ_x^2 will vary, since it depends on the procedure used in collecting the data of the study. In words, σ_x^2 means the variance of the obtained scores of all of the teachers in the population about their own mean. Obviously, what this variance is depends on how the obtained scores are obtained. It will, for example, be greater if different classes are visited by different observers than if all are visited by the same observers (pp. 309-310).

It should be noted that, according to Medley and Mitzel, true score is defined as the expected mean score in a variety of circumstances, but the observed score variance is computed on scores collected under a set of circumstances which may be quite different. An investigator would need to determine the range of circumstances over which true score variance should be estimated, and no clear basis is given for this decision. Such uncertainty may perhaps best be overcome by setting aside the notion of true score (with its associations of absoluteness) and using instead, the concept of a "universe score", as defined by Cronbach et al. (1972):

> The score on which the decision is to be based is only one of many scores that might serve the same purpose. The decision maker is almost never interested in the response given to the particular stimulus objects or questions, to the particular tester, at the particular moment of testing. Some, at least, of these conditions of measurement could be altered without making the score any less acceptable to the decision maker. That is to say, there is a universe of observations, any of which would have yielded a usable basis for the decision. The ideal datum on which to base the decision would be something like the person's mean score over all acceptable observations, which we shall call his "universe score." The investigator uses the observed score or some function of it as if it were the universe score. That is, he generalizes from sample to universe. The question of "reliability" thus resolves into a question of accuracy of generalization, or generalizability.
>
> The universe of interest to the decision maker is defined when he tells us what observations would be equally acceptable for his purpose (i.e.,

> would "give him the same information"). He must describe the acceptable set of observations in terms of the allowable conditions of measurement. This gives an operational definition of the class of procedures to be considered (p. 15).

Given this notion of universe score, we may think of reliability, or generalizability, as the expected value of the correlation between the set of observed scores and other sets of observed scores from the universe of interest. Thus the onus is on the investigator to define his universe of interest. This is as it should be; the universe to which a set of observations are to be generalized will depend on the practical or theoretical purpose of the investigator. Normally one would expect that set of observations will be chosen in such a way as to ensure that they are representative of this universe; certainly the way in which the observations are obtained places limits to the universe to which they can be generalized.

Suppose, as an example, that (using the notation of Medley and Mitzel, 1963), r observers have visited c classrooms (or teachers) in s situations. The investigator should specify how the r observers were chosen and trained; how the c classrooms were selected, and in what way the situations (times of the visits) were chosen (random selection, availability of transport, etc). Then the universe of generalization is that of observations collected in the same way. An investigator might wish to narrow his universe of generalizability, perhaps by considering only the subset of observations collected by observer A; in this case he might expect to find a higher coefficient of generalizability, but this is obtained at the expense of having a universe of generalization which is (presumably) of less interest.

Estimating reliability

Suppose that n visits are made to each of t teachers. On each visit, an estimate is made of some characteristic X of the teacher. (It is assumed, although it is not necessary to the development which follows, that X is estimated by systematic observation.) Each teacher has a universe score on characteristic X, where the universe is defined by the way in which the observations are made. The universe score will be estimated by the average of the n estimates obtained for each teacher, and it is desired to find an appropriate coefficient of reliability (or generalizability) for this average score.

The data might be displayed in a matrix such as that shown in Table 1. In this matrix, x_{ij} is the score awarded on the ith visit to the jth teacher. The visits are regarded as equivalent to one another, and no distinction is intended between visits (in the sense that there is no correspondence between, say, visit 5 to one teacher, and visit 5 to another). Application of the standard one-way analysis of variance to the data in Table 1, with teachers being the only factor, and visits treated as replications, would yield the usual ANOVA summary table of the form shown in Table 2. In this table, σ_w^2 is the variance of scores attributed to a single teacher (averaged over teachers); σ_t^2 is the variance of the teachers' universe scores. These can be estimated from the sample values:

$$\hat{\sigma}_w^2 = MS_w$$

and

$$\hat{\sigma}_t^2 = (MS_t - MS_w)/n \quad ,$$

TABLE 1

DATA MATRIX FOR THE CASE WHERE EACH TEACHER IS OBSERVED ON n OCCASIONS

		TEACHERS								
		1	2	3	4	.	.	j	.	t
	1	x_{11}	x_{12}	x_{13}	.	.	.	x_{1j}	.	x_{1t}
	2	x_{21}	x_{22}	.	.	.				
O	3	x_{31}	x_{32}	.	.					
C	.	.	.	.						
A	.	.	.							
S	.	.								
I	.	.								
O	.	.								
N	.	.								
S	i	.						x_{ij}		
	.	.								
	.	.								
	n	x_{n1}	.	.						x_{nt}

TABLE 2

ANALYSIS OF VARIANCE SUMMARY TABLE FOR n OBSERVATION PERIODS ON EACH OF t TEACHERS

Source of Variation	Degrees of Freedom	Mean Square	Expected Mean Square
Among teachers	$t - 1$	MS_t	$\sigma_w^2 + n\sigma_t^2$
Within teachers	$t(n - 1)$	MS_w	σ_w^2
Total	$nt - 1$		

where $\hat{\sigma}_w^2$ and $\hat{\sigma}_\dagger^2$ are unbiased estimates of σ_w^2 and $\sigma_\dagger^2$. Since $\sigma_\dagger^2$ is the variance of universe scores, corresponding to "true" variance in the traditional approach, and $(\sigma_\dagger^2 + \sigma_w^2)$ is the observed variance, the reliability of an individual score will be given by:

$$\rho_{11} = \sigma_\dagger^2/(\sigma_\dagger^2 + \sigma_w^2) \quad ,$$

and is estimated by

$$r_{11} = \hat{\sigma}_\dagger^2/(\hat{\sigma}_\dagger^2 + \hat{\sigma}_w^2) \quad .$$

Substituting the known expressions for $\hat{\sigma}_\dagger^2$ and $\hat{\sigma}_w^2$, we obtain:

$$r_{11} = \frac{MS_\dagger - MS_w}{MS_\dagger + (n - 1)MS_w} \quad ,$$

which estimates the reliability of a single score. This would probably be of only mild interest, since the mean of the n scores awarded to each teacher is the estimate of universe score which would normally be used. We may note that the reliability of this mean score is the same as that of the sum of the n scores contributing to it, and hence may be obtained by the application of the generalized Spearman-Brown formula:

$$\rho_{nn} = \frac{n\rho_{11}}{1 + (n - 1)\rho_{11}} \quad .$$

Substitution of the previously obtained estimate for ρ_{11} in this formula yields the estimate:

$$r_{nn} = \frac{MS_\dagger - MS_w}{MS_\dagger} \quad .$$

The development above is similar to that provided by Ebel (1951) for the analysis of the reliability of ratings, and the coefficient r_{11}

is formally identical (Fisher, 1925, pp. 222-224; Ebel, 1951, pp. 409-410; Haggard, 1958, pp. 10-12) to the intraclass correlation coefficient. Thus the correlational notion of reliability and the variance ratio approach are again seen to coincide.

Typically, observational data are expensive and time-consuming to collect. Observation schedules can be disrupted by bad weather conditions, unanticipated holidays, excursions, illnesses, transport problems, and a host of other causes. Even the best of intentions cannot ensure that equal numbers of visits are made to all teachers, and an investigator may decide to proceed with the data he has, rather than to spend further time and money rescheduling visits to make up for those missed. How can reliability be estimated under these circumstances?

Clearly, such a situation would raise no conceptual difficulties if we were to construct a symmetrical correlation table, in the manner of Fisher (1925, pp. 213-214), and compute the standard interclass correlation from this table. Of course, the number of entries in the table would likely be extremely large. However, the analysis of variance approach can again be used to make the computation much simpler. The major difference in this case is that we can no longer use n (the number of visits per teacher) in the computations, but instead must use an "averaged" value of n given by:

$$n_o = \frac{1}{t - 1} \left\{ \sum_{j=1}^{t} n_j - \sum_{j=1}^{t} n_j^2 \Big/ \sum_{j=1}^{t} n_j \right\} .$$

(See Snedecor, 1946, p. 234; Ebel, 1951, p. 413; or Haggard, 1958, p. 14). The intraclass correlation (reliability of a single score) would then be given by:

$$r_{11} = \frac{MS_t - MS_w}{MS_t + (n_o - 1)MS_w}$$

We may note that the formula for computing r_{nn} (the reliability of the mean score for each teacher) is unchanged, since it does not contain n. A well-designed study would probably aim to have n_j's which are as nearly equal as possible, so that, in practice, the values being "averaged" would not be widely different.

For most purposes, estimates of reliability obtained in this way would probably be quite adequate. For those users wishing to provide information about the precision of these estimates, Ebel (1951, pp. 413-414) and Haggard (1958, pp. 22-25) have described a method by which confidence intervals can be constructed around the reliability estimate. Interested users should consult either of the above references for details.

As an illustration of the computational procedures involved in the estimation of reliability, consider the raw data presented in Table 3. These data were obtained by John Herbert and his associates, in an unpublished study, conducted at The Ontario Institute for Studies in Education. Thirty teachers were observed, and their behaviors coded once per minute, using a modified version (SAL II) of the System for Analyzing Lessons (Herbert, 1967). Observation periods of 50 minutes were chosen, the number of such periods available for each teacher ranging from five to nine, and totalling 187 for the 30 teachers. Standard one-way analysis of variance with unequal n's yielded the summary displayed in Table 4. The same table contains also the calculations necessary to obtain the reliability of the mean scores, and also of the single scores.

From this analysis, we may observe that the reliability of scores obtained from single observation periods is relatively

TABLE 3

RAW DATA: FREQUENCY OF OCCURRENCE OF TEACHERS' "PRESENTING" BEHAVIOR, BASED ON ALL AVAILABLE OBSERVATION PERIODS OF 50 MINUTES

Teacher	Observations	n	Mean	St. Dev.
1	14,28,18,10,20	5	18.0	6.78
2	15,16,22,19,20,24,21,14,19	9	18.9	3.33
3	28,18,12,13,18,22, 7	7	16.9	6.94
4	19,25,23,34,20,31,27,25	8	25.5	5.13
5	21,21,18,24,18	5	20.4	2.51
6	21,27,26,14,30,27	6	24.2	5.77
7	29,17,28,16,32,29,25	7	25.1	6.25
8	23,34,19,17,26,31,27,23	8	25.0	5.73
9	24,13,25,33, 8	5	20.6	10.01
10	24,25,19,15,19,10	6	18.7	5.61
11	22,19,24,23,19,18	6	20.8	2.48
12	20,13,24,21,25,26,19	7	21.1	4.45
13	27,10,27,18,21	5	20.6	7.09
14	24,34,31,22,23,30,30,37	8	28.9	5.41
15	20,24,35,22,38,36	6	29.2	8.01
16	22,23,16,22,25	5	21.6	3.36
17	15,17, 5,13,14,16	6	13.3	4.32
18	32,23,26,19,25	5	25.0	4.74
19	33,26,40,37,33,33,26	7	32.6	5.19
20	13, 8,15,13,23	5	14.4	5.46
21	17,25,19,15,30,13,19	7	19.7	5.91
22	19,12,25,13,14	5	16.6	5.41
23	29,27,27,24,18,31,33,20	8	26.1	5.19
24	20,32,28,29,23	5	26.4	4.83
25	15,16,18, 9,12,14	6	14.0	3.16
26	27,36,23,28,32,23,36	7	29.3	5.53
27	28,22,19,26,23	5	23.6	3.51
28	12,23,21,18,17	5	18.2	4.21
29	16, 9,23,25,25,29,20	7	21.0	6.71
30	16,15,25,26,30,22	6	22.3	5.89

TABLE 4

ANALYSIS OF VARIANCE AND RELIABILITY COMPUTATIONS BASED ON UNEQUAL NUMBERS OF 50-MINUTE VISITS TO EACH OF 30 TEACHERS

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
Among teachers	4239.625	29	146.19	4.83
Within teachers	4753.000	157	30.27	
Total	8992.625	186		

$$\sum_{j=1}^{30} n_j = 187 \qquad \sum_{j=1}^{30} n_j^2 = 1207$$

$$n_0 = \frac{1}{29}\left(187 - \frac{1207}{187}\right)$$

$$= 6.226$$

Reliability of a single score

$$r_{11} = \frac{146.19 - 30.27}{146.19 + (5.226)(30.27)}$$

$$= .381$$

Reliability of a mean score

$$r_{nn} = \frac{146.19 - 30.27}{146.19}$$

$$= .793$$

low (.381), and indicates that such scores would not, in themselves, be very useful for the detection of relationships with other variables. However, the mean scores for each teacher are considerably more reliable (.793), indicating that these scores could correlate substantially with other variables, if the underlying relationships were strong enough.

The universe of generalizability

The procedures described in preceding sections are intended to be used with sets of observations collected for purposes other than just the estimation of reliability. An investigator interested in finding correlates of pupil growth, for instance, can use these procedures to estimate the reliability of the observational measures that he uses, the only requirement being that the best estimate of a universe score must be the mean of the observed scores for that individual. Unlike the approach of McGaw et al. (1972), this treatment yields a single coefficient. Consideration of the nature of the two approaches will make clear why this is so.

McGaw et al., like Medley and Mitzel, are concerned with the analysis of a reliability study; each require that data be collected under circumstances which provide for the controlled variation of certain factors, or facets (teacher, observer, situation), so that variation in the measure obtained can be attributed to each of these sources. The coefficients produced describe the generalizability to different universes, the universes being defined by the controlling of one or more facets.

The present approach yields a single coefficient which describes the generalizability of the scores to a universe of which the current observations are assumed to be representative. Thus the reporting of a

coefficient alone would be inadequate without a description of the universe to which generalization is intended. It would normally be the responsibility of the investigator to demonstrate that the observations he has obtained are indeed representative of the universe to which he claims to generalize. This should not be thought of as an imposition; in fact it ought to be regarded as essential if any generalization at all is to be made from the study.

Definition of the universe of generalization should not be a difficult task if the investigator is guided by the description given by Cronbach et al. (1972): "The universe of interest to the decision maker is defined when he tells us what observations would be equally acceptable for his purpose (p. 15)." Suppose, for instance, that a single observer had made three visits to each of, say, ten teachers. The investigator probably has no special interest in those particular teachers, and, presumably, any other ten teachers selected according to the same criteria would have been equally acceptable. Similarly, he would probably have been just as satisfied with any other observer having similar training and skills. He might then describe the universe of generalization as the universe of observations made in three visits, by a single observer, similarly trained and competent, to groups of ten teachers selected in the same way. If all observations had been made during the first period of the day, or while History was being taught, then the universe of generalization would be narrowed accordingly. If a team of observers had been used in place of the single observer, then the universe would be broadened.

The last instance provides an example of a situation which might appear paradoxical when viewed in the traditional way, but presents no

problems when described in terms of the concepts of generalizability theory. If we accept that there are likely to be systematic differences between observers in the scores that they award, then it follows that the "error" variation will be greater with a team of observers than if a single observer had been used. Since it follows that we will get greater reliability using a single observer, can we then conclude that this procedure is preferable? In traditional terms, we would probably answer "no", on the grounds that the increase in reliability is almost certain to be accompanied by a decrease in validity. However, with neither a suitable definition of validity, nor a means of measuring it, such an argument is not easy to sustain. In terms of generalizability theory, a much simpler and more satisfying explanation is possible: it is true that the coefficient of generalizability is higher in the case of a single observer, but the universe to which generalization can be made is of considerably less interest. Such an example provides a clear illustration of the importance which ought to be attached to a clear statement by the investigator of his intended universe of generalization.

Empirical investigations

1. Reliability and frequency of occurrence

The techniques outlined previously were applied to the Herbert data, in order to estimate the reliabilities of measures of the relative frequencies of the SAL II categories. The number of measures which could possibly be extracted from such a data set is limited only by the researcher's imagination, and the 20 presented in Table 5 are not claimed to be representative; they are, however, amongst the more obvious measures one

TABLE 5

RELIABILITIES OF MEASURES OF FREQUENCIES OF SAL II CATEGORIES, BASED ON 50-MINUTE OBSERVATION PERIODS

Variable number	SAL II Category	Percent of observations (mean)	Intraclass correlation r_{11}	Reliab. of mean score r_{nn}
	Social Orientation (S)			
1	1. Pupil	54.0	.357	.775
2	2. Nobody	13.0	.259	.685
3	3. Whole class	20.0	.338	.760
4	4. Group	9.3	.253	.678
	Nature of Interaction (I)			
5	1. Presenting	44.4	.381	.793
6	2. Interacting equally	0.2	.064*	
7	3. Calling on	14.9	.134	.491
8	4. No interaction	12.9	.259	.685
9	5. Watching, listening	24.0	.180	.578
10	6. Responding	1.4	-.003*	
	Affect Tone (A)			
11	1. Curt, angry	1.2	.444	.833
12	2. Unhappy	0.03	-.030*	
13	3. Neutral	90.1	.474	.849
14	4. Happy, pleasant	5.1	.490	.857
15	5. Supportive	1.5	.298	.726
	Subject Matter (M)			
16	1. Conduct	2.6	.259	.685
17	2. Lesson subject	83.1	.088*	
18	3. No subject activity	2.1	.146	.516
19	4. Another school subject	0.6	.053*	
20	5. Routine	9.1	-.016*	

* Intraclass correlation not significantly different from zero ($p > .01$). Reliability of the mean score was not computed.

might be interested in computing. For each of the 20 variables, the analysis was identical to that carried out for variable 5 in Table 4, except that where the intraclass correlation proved not to be significantly different from zero[1] ($p>.01$), the reliability of the mean score was not computed.

Examination of Table 5 may serve to confirm many of the reader's expectations. Reliabilities of measures based on single visits (r_{11}) are all quite low (below 0.4, in all except three cases); but, for some categories, the reliabilities of measures based on 6-7 visits (r_{nn}) are substantial, exceeding 0.8 in some cases. A quick examination of Table 5 might lead one to the conclusion that the unreliable measures were those with the smallest frequencies, probably in accord with one's expectations. However, such generalization would be unwarranted, as the exceptions are, in many ways, more interesting than those which fit the pattern.

Variables 6 (Interacting equally), 10 (Responding), 12 (Unhappy), and 19 (Another School Subject) all had near-zero reliabilities, and in each case, we may guess that their frequencies of occurrence were too small to make reliable measurement possible. It is interesting to note, though, that two categories of the dimension Subject Matter (17: Lesson Subject, and 20: Routine) occurred with substantial frequencies (83.1 and 9.1 percent, respectively, of observations), but yielded

---

[1] A significant intraclass correlation is obtained when the analysis of variance from which the intraclass correlation is obtained yields a significant F-ratio (see Haggard, 1958, pp. 19-22).

reliabilities extremely close to zero. Clearly, a high frequency of occurrence does not guarantee reliability, and in these two cases, we have to conclude that, while there were substantial differences over occasions, these differences were not maintained consistently amongst teachers.

It is certainly interesting, and perhaps important, to note the relatively high reliabilities recorded for most of the categories of the dimension "Affect Tone". Concern has been voiced by the developer of SAL II (Herbert, personal communication) over the fact that very few observations were recorded in categories other than 3 (Neutral), and that the frequencies in the other (more interesting?) categories may have been too small to have yielded any useful information. However, the high reliabilities recorded for categories 1 (Curt, angry), 2 (Happy, pleasant) and (to a lesser extent) 5 (Supportive) demonstrate that it is at least possible to obtain reliable measures from behaviours which occur quite infrequently. Considering, for instance, category 1 (Curt, angry), we would have to conclude that, in spite of its relatively rare occurrence (1.2 percent, on average), teachers do show some consistency in the extent to which they display, or do not display, anger in their teaching. It seems at least reasonable to imagine that whatever the effects of anger might be, they would occur with relatively few displays of anger on the part of the teacher. Consequently, a researcher interested in studying the incidence and/or effects of the use of anger in teaching need not be deterrred by its relatively infrequent occurrence in natural settings.

Conclusions of a general nature can at best be cautionary, rather than prescriptive. Certainly, if a particular behavior is of sufficient interest, we should not be deterred from attempting to measure it solely

on the grounds that its occurrence is relatively infrequent. Nor, on the other hand, can we assume that the accumulation of large numbers of observations of a particular behavior provides some kind of guarantee that we have achieved precision of measurement. What really matters is not the number of times that the particular behavior has been observed, but whether the subjects of the observation have differed consistently in the extent to which they display that behavior. This cannot be inferred from considerations of frequency alone, but needs to be determined by an analysis of the type described in earlier sections of this paper.

2. The effect of varying the number and length of the observation periods

For observation periods of given length, it is well known that the reliability increases as the number of observation periods is increased, and the relationship can be described by the familiar Spearman-Brown formula. For a fixed number of observation periods, it seems at least intuitively reasonable to expect that the reliability would be greater for observation periods of greater length. The precise nature of the relationship between reliability and length of observation period has not previously been explored. This is an empirical, rather than a theoretical investigation, since no theory presently exists to predict the nature of the relationship.

Using the same body of data, we may ask what reliability would have been obtained had each observation period been cut short after 10, 20, 30, or 40 minutes, rather than the 50 minutes that we have been using. (The analysis could also be extended to periods greater than 50 minutes, too, except that the number of periods available, and hence the precision of the reliability estimates, decrease quite drastically, even for 60 munute periods.) The analysis is presented in detail only for

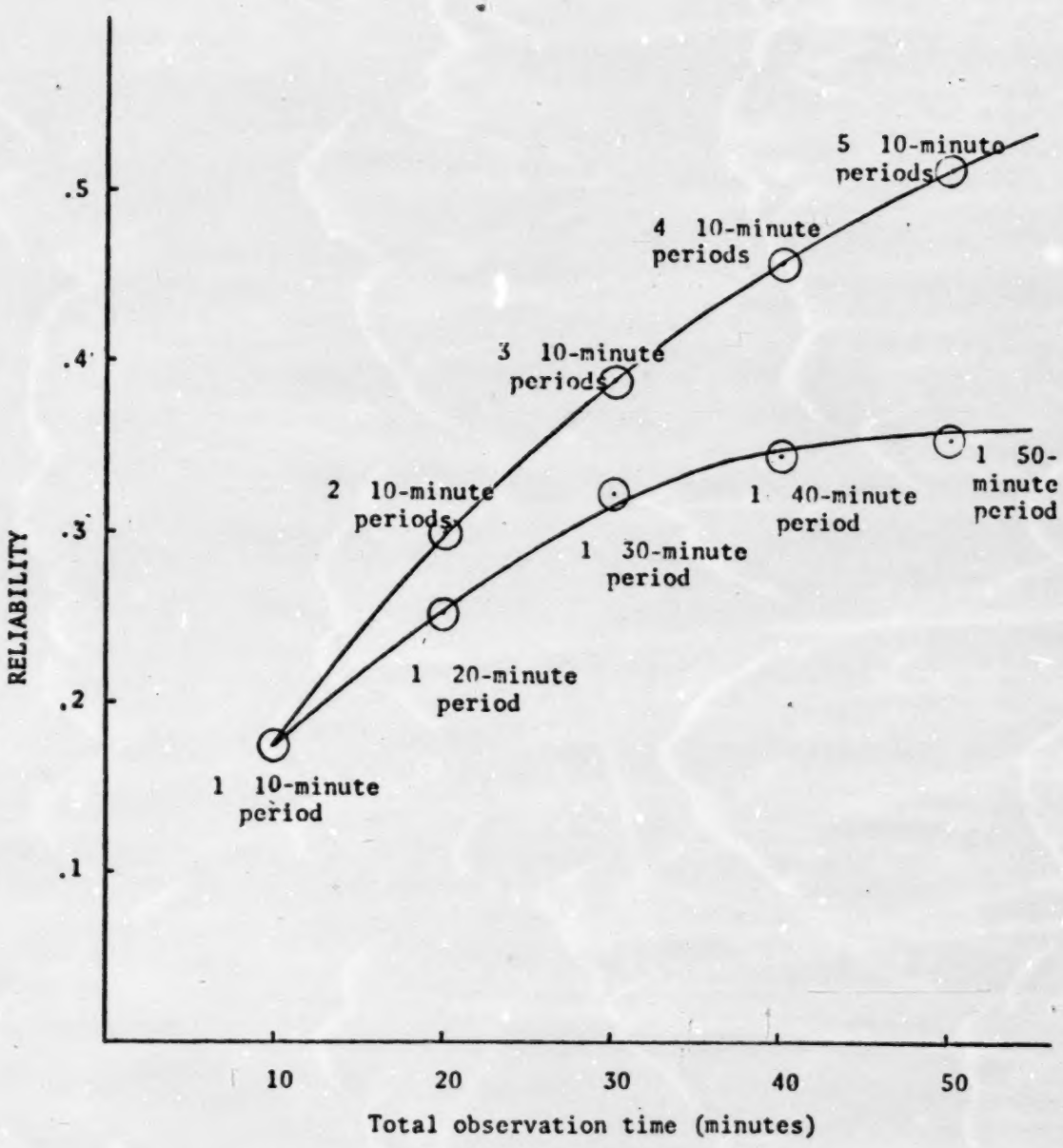

FIGURE 1. The effect on reliability of increasing both the number and the length of the observation periods: variable 1.

variable 1: the percentage of the teachers' time spent relating to individual pupils.

We recall that reliability measures the extent to which our set of observations is representative of the wider universe of admissible observations (Cronbach et al., 1972, p. 20). It seems reasonable, then, to expect that increasing the number of independent samples of behavior would be more effective in producing a representative set of observations than taking the same set of occasions, and observing them for longer periods of time. In the latter case, there is the possibility that the extra observations may contribute very little to the representativeness of the data, particularly if the behavior being observed differs little from that which preceded it.

This expectation proves to be justified, and is illustrated in a specific case by Figure 1. For variable 1, the reliability of scores based upon one ten-minute observation period per teacher was estimated to be 0.176. The reliabilites obtainable from two, three, four, and five such observation periods are found by the use of the Spearman-Brown formula; the reliabilities obtained from single observation periods of 10, 20, 30, 40, and 50 minutes are determined empirically. Figure 1 provides a clear demonstration of the general result that, for fixed total observation time, greater reliability is achieved by the use of a larger number of shorter, independent observation periods.

This demonstration is, however, dependent on the accuracy of one estimate -- the reliability of a ten-minute observation period -- and this is the least precise of the measurements being used. Table 6, and Figure 2, which is derived from it, illustrate the same general principle in a way which is not so heavily dependent upon the accuracy of that one figure.

TABLE 6

THE RELIABILITIES OF MEASURES OF VARIABLE 1, AS BOTH THE NUMBER AND THE LENGTH OF THE OBSERVATION PERIODS ARE VARIED

Number of observation periods	Length of each observation period (in minutes)				
	10	20	30	40	50
1	.176	.252	.323	.346	.357
2	.299	.402	.488	.514	.526
3	.391	.502	.589	.614	.624
4	.461	.574	.656	.679	.689
5	.516	.627	.705	.726	.735
6	.562	.669	.741	.761	.769

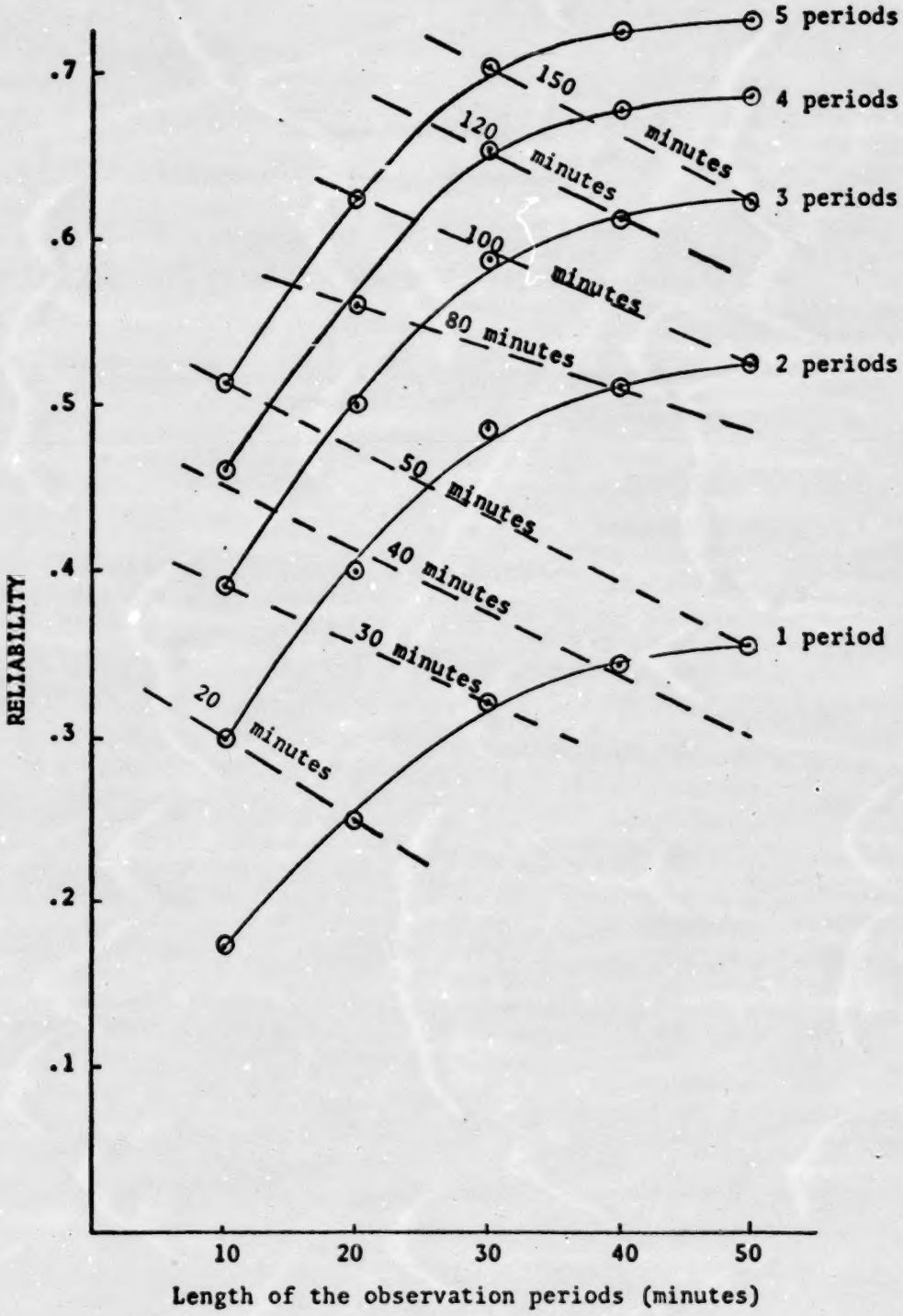

FIGURE 2. The reliabilities of measures of variable 1, as both the number and the length of the observation periods are varied.

From Table 6, for example, we may observe that the reliability of 0.176 obtained from one ten-minute visit could be increased to 0.357 by increasing the length of the visits by a factor of five, but to 0.516 by making five times as many visits. In Figure 2, the solid curves show the increase in reliability with the lengths of the observation periods for 1, 2, 3, 4, and 5 observation periods, while the broken lines join points which correspond to equal total observation times. The downward slopes of these lines demonstrate the decrease in reliability as the observation periods are made longer and fewer.

These results, which were entirely predictable, are ones to which more heed might be paid in the planning of observational studies. Periods of observation are frequently very long, running, for instance, over two hours in the Herbert study. Given the high costs of transportation, the desire to achieve the maximum observation time for each mile travelled is understandable. However, it does appear that two hours of consecutive observations would be, at best, only marginally more representative (and hence more reliable) than one hour, or perhaps even half an hour. Had the observation periods been shorter, and more numerous, it seems at least likely that the observation schedule could have been arranged in such a way as to achieve substantially greater reliability, perhaps even at lower cost.

Conclusion

This study has provided a rationale for, and a relatively simple method of, estimating the reliability of an observational measure. The method described does not require the setting up of a separate "reliability study", but may be employed in examining data which has been collected for correlational or experimental purposes. It does not require the use of multiple observers, although it is entirely appropriate in the situation where multiple observers are used. Interpretation of the coefficient produced is straightforward, provided the conditions of observation are well described.

The empirical findings reported in the study establish clearly that high frequencies of occurrence are not necessary pre-requisites for the reliable measurement of behavior. Further, it has been demonstrated that reliability increases quite regularly as both the number and the length of the observation periods are increased, but that, for fixed total observation time, higher reliability is achieved by the use of a larger number of shorter observation periods. These findings may be of assistance to a researcher in the planning stage, who wishes to design his study in such a way as to maximize the reliability of the measures he obtains, and so increase the probability that he will find what he is looking for.

REFERENCES

Brown, B. B., Mendenhall, W., & Beaver, R. The reliability of observations of teachers' classroom behavior. Journal of Experimental Education, 1968, 36, 1-10.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. The dependability of behavioral measurements: Generalizability of scores and profiles. New York: Wiley, 1972.

Ebel, R. L. Estimation of the reliability of ratings. Psychometrika, 1951, 16, 407-424.

Fisher, R. A. Statistical methods for research workers. Edinburgh: Oliver & Boyd, 1925. Thirteenth edition, reprinted 1967.

Flanders, N. A. Analyzing teaching behavior. Reading, Mass.: Addison-Wesley, 1970.

Frick, T., & Semmel, M. I. Observational records: observer agreement and reliabilities. Paper presented at the 1974 meeting of the American Educational Research Association, Chicago, April 16, 1974.

Haggard, E. A. Intraclass correlation and the analysis of variance. New York: Dryden Press, 1958.

Herbert, J. D. A system for analyzing lessons. New York: Teachers' College Press, 1967.

Lord, F. M., & Novick, M. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

McGaw, B., Wardrop, J. L. & Bunda, M. A. Classroom observation schemes - where are the errors? American Education Research Journal, 1972 9, 13-27.

Medley, D. M., & Mitzel, H. Application of analysis of variance to the estimation of the reliability of observations of teachers' classroom behavior. Journal of Experimental Education, 1958, 27, 23-35.

Medley, D. M., & Mitzel, H. Measuring classroom behavior by systematic observation. In N. L. Gage (ed.) Handbook of Research on Teaching, Chicago: Rand-McNally, 1963, Pp. 247-328.

Rosenshine, B. Teaching behaviors and student achievement. Windsor, Berks.: National Foundation for Educational Research in England and Wales, 1971.

Rosenshine, B., & Furst, N. F. The use of direct observation to study teaching. In R. M. W. Travers (ed.) Second Handbook of Research on Teaching. Chicago: Rand-McNally, 1973, Pp. 122-183.

Snedecor, G. W. Statistical Methods. (4th ed.) Ames, Iowa: State College Press, 1946.

Stanley, J. C. Reliability. In Thorndike, R. L. (ed.) Educational Measurement. Washington: American Council on Education, 1971.

Westbury, I. The reliability of measures of classroom behavior. Ontario Journal of Educational Research, 1967, 10, 125-138.